The National Geospatial Digital Archives—Collection Development: Lessons Learned

Tracey Erwin, Julie Sweetkind-Singer, and Mary Lynette Larsgaard

Abstract

There are many similarities between building a geospatial digital archive and building a hard-copy map collection, and two major ones are the necessity to have a collection development policy and the amount of hard work required to seek out and acquire the resources. Two institutions, University of California at Santa Barbara and Stanford University, the initial partners in the National Geospatial Digital Archives (NGDA), chose to collect digital data that was in line with each library's standard collection strengths and responsibilities. Collection development policies were written for the project as a whole and for each partner institution. While based on traditional paper map policies, these geospatial collection development policies are tailored specifically for digital data by including sections on metadata, versioning, file formats, proprietary formats, data set size, and ownership/access considerations.

During the acquisition phase of the contract a considerable amount has been learned about file formats, data acquisition of compressed vs. uncompressed files, short-term storage prior to repository ingest, and metadata creation. While metadata creation at the collection-level/series-level has been relatively easy the acquisition phase has underscored the challenges inherent in creating accurate item-level metadata. One of the central findings of the NGDA experience is that format information is vital for long-term preservation. Thus, the need to understand file formats and specifications has led to the creation of a format registry specifically for geospatial materials.

LIBRARY TRENDS, Vol. 57, No. 3, Winter 2009 ("The Library of Congress National Digital Information Infrastructure and Preservation Program," edited by Patricia Cruse and Beth Sandore), pp. 490–515

Introduction

The Library of Congress's National Digital Information Infrastructure Preservation Program's (NDIIPP) ambitious goal of collecting and preserving our nation's digital heritage has been under way for over four years. Of the eight initial cooperative agreements, two were focused on geospatial data and imagery. The North Carolina Geospatial Data Archiving Project has pursued preservation of state and local digital geospatial data. The goal of creating the National Geospatial Digital Archive has been to build a collecting network for the archiving of geospatial images and data. The stated goals of the NGDA are to:

- create a new national federated network committed to archiving geospatial imagery and data;
- investigate the proper and optimal roles of such a federated archive, with consideration of distant (dark) backup and migration;
- collect and archive major segments of at-risk digital geospatial data and images;
- develop best practices for the presentation of archived digital geospatial data;
- develop partner communication mechanisms for the project then and ongoing;
- develop a series of policy agreements governing retention, rights management, obligations of partners, interoperability of systems, and exchange of digital objects.

Geospatial data represent a vast array of knowledge about the earth ranging from highly accurate information about elevation, land cover, glacial extent, sea surface temperatures, air and water quality, to road and other infrastructure networks (Figure 1).

Geospatial datasets combined with visualization software allow for complex analysis of data about place. Many disciplines use geospatial data to discern what is happening on our planet and to predict possible future scenarios. To collect and archive a part of this rich data is the mandate of the National Geospatial Digital Archive (NGDA). This mandate fits well with the explicit role of academic libraries, which has long been to preserve the intellectual record of the past as well as to capture the output of contemporary researchers, scientists, and scholars. The Stanford libraries hold several editions of Charles Darwin's *On the Origin of Species by Means of Natural Selection,* and numerous works of the noted geologist Charles Lyell. These seminal works greatly influenced how the world was viewed in the late nineteenth century. As such, they are critical works for an academic library to hold.

Similarly, the choice of the NGDA to archive the California Geological Survey (CGS) is critical to the mission of the University of California, Santa Barbara and Stanford University today. Geological mapping of Cali-

Figure 1. Digital Raster Graphic of San Francisco

fornia at the most detailed scale of 1:24,000 is *still* incomplete in 2008. In a state known for its seismic activity this mapping is of great importance. Capturing the ongoing output of the CGS will provide benchmark data for future generations of researchers. Preserving such data is a prudent step in a world of ever tightening state budgets and ongoing cuts to the agencies who produce the data and who do not have a mandate to pre-

serve, only to produce geological surveying. Archiving the CGS data will provide potential access to the widest possible segment of the public.

In some regards all collection development is an educated guess regarding what will be valued later. How data will be important to future generations is not always knowable. One such example is the 112-year-old National Weather Service collecting station at Mohonk, New York. The consistency of the data (only five people have recorded the data in the entire history of the station) has yielded a remarkable record of climate change as measured by weather data and phenological observations (DePalma, 2008).

Although many federal agencies that produce and distribute geospatial data may be presumed to be engaged in preservation efforts, there is a difference between saving bits and truly preserving them. Also, it seems prudent to preserve the data that are most important to our constituents. Even if some other entity may also be preserving the data, redundancy is a failsafe in itself. Also, priorities change: imagery that today may seem destined for preservation by national agencies may abruptly cease to be collected. Finally, there is no choice between whether to collect digital or analog data. All data are already digital or are being digitized.

The University of California, Santa Barbara (UCSB), in partnership with Stanford University, was awarded an NDIIPP contract in October 2004. As noted earlier, the mission of the NGDA is to "collect and archive major segments of at-risk digital geospatial data and images." The breadth and quantity of geospatial data and imagery available today necessitates judicious selection. Remote sensing image collections *already* held by forty-eight academic libraries and the National Archives and Records Administration are vast: these collections range from ten thousand to eighteen million images (Thiry, 2006 pp. 312–313). With remote sensing imagery such as MODIS being produced at the rate of one terabyte a day, the potential for data archiving is increasing exponentially (Leptoukh et. al., 2001).

With this output volume in mind, and from early conversations about collecting, it was clear that formal collection development policies would be vital to guide the process. With a few key collections in mind at the project's inception, work began early in the contract period to create collection development policies that would address both the national nature of the NGDA, and the institutional realities of the two universities hosting the archive. This resulted in the creation of three collection development policies. The first is an overarching policy designed to guide any collecting body that signed onto the NGDA as a node. The second and third are specific policies written by the founding nodes, UCSB and Stanford.

Much of the context in the discussion of collection development policies involves the concept of risk. All collecting at cultural institutions is done to ensure the persistence over time of knowledge that has been

deemed valuable. While much of the work of the NGDA has focused on obtaining large sets of data and imagery, many of which have been produced by the federal government, it is important to remember the output of smaller producers. Perhaps the data most at risk are those created by institutes, individuals, and small research centers that lack institutional resources to do their own archiving. Such entities will need to be approached thoughtfully to discuss the need for long term preservation of their work and to decide if a large university archive is the proper place to hold their content.

The Collection Development Policies

The first policy to be written governs the collection efforts of the NGDA as a whole. It is intended to serve as both a collection development policy (CDP) in its own right for all of the NGDA nodes and as a model for institutions who wish to develop their own geospatial collection development policy. At the beginning of the contract period approximately eight academic or public map libraries were surveyed about their geospatial collections in order to discover what kind of collecting policies were in place for such materials. A number of libraries indicated that although their geospatial holdings were increasing, it was in an ad hoc manner, without a systematic collecting policy. The NGDA policy is included as attachment 1 following this article, while the individual policies for UCSB and Stanford can be found on the NGDA website at http://www.ngda.org/.

Digital geospatial collection activities at the two NGDA partner institutions follow the research interests and institutional needs of each respective institution. The amount of geospatial data and imagery produced in the United States each year is vast, requiring both institutions to make judicious choices about what to collect and the frequency with which it will be captured. These institution-specific CDPs share some features and depart in some areas from the traditional paper collecting policies upon which they are based.

The major similarities between each institution's digital geospatial collection development policy and the collection development policy for its paper collection are in the areas of subject, scope, and region of coverage. While the overarching policy suggests certain parameters, each institution interprets the general guidelines differently to fit local needs.

Each institution's geospatial CDP parallels the traditional focus of the respective university. UCSB's collection has always focused on southern California, and includes its prominent aerial photography and imagery collection. Stanford has always had an emphasis on northern California, the San Francisco Bay Area, and the Monterey Bay. Table 1 illustrates some of the notable similarities and differences in collecting scope at UCSB and Stanford University.

UCSB —Alexandria Digital Library CDP	STANFORD—Branner Earth Sciences Library and Map Collection CDP
A. Geography Primary collecting emphasis is on the geographic area firstly of Santa Barbara County, and of California as a whole, and secondarily of the United States.	A. Geography Primary collecting emphasis is on the geographic area firstly of Santa Clara and San Mateo counties, and of California as a whole, and secondarily of the United States.
B. Subject Potential types of materials are split across physical and human/cultural geography. The list has been derived from the Library of Congress G Schedule and then modified.	B. Subject Potential types of materials are split across physical and human/cultural geography. The list has been derived from the Library of Congress G Schedule and then modified.
• For Santa Barbara County, UCSB will collect all subjects listed. • For California as a whole, UCSB will concentrate on the following subjects: topography biogeography exception: agriculture, which is UC Davis' purview within the UC Libraries system historical geography • Map-format materials (e.g., maps, diagrams, sections, views, profiles)	• For Santa Clara and San Mateo Counties, Branner will collect all subjects listed. • For California as a whole, Branner will concentrate on the following subjects: topography geology environmental aspects of such areas as oceanography, climatology, economics, etc. • Map-format materials (e.g., maps, diagrams, sections, views, profiles)
a. Physical geography - Mathematical geography (surveying and cartography, etc.) - Physiography (e.g., topography, bathymetry, and hydrography including nautical charts) - Hydrology - Geology, geophysics, mineral resources, and soils - Climatology - Biogeography (land use/land cover)	*a. Physical geography* - Mathematical geography (surveying and cartography, etc.) - Physiography (e.g., topography, bathymetry, and hydrography including nautical charts) - Hydrology - Geology, geophysics, mineral resources, and soils - Climatology - Biogeography (land use/land cover)
b. Human and cultural geography - Political geography - Public lands, ethnic reservations - Demography, census - Transportation and communication - Historical geography	*b. Human and cultural geography* - Political geography - Public lands, ethnic reservations - Demography, census - Transportation and communication - Historical geography

Table 1: Similarities and Differences between Policies

Geospatial collection policies for both universities differ from their paper policy counterparts in many ways. First, the potential patrons extend far beyond the usual university scope of students, faculty, and staff to include all citizens of the United States, present and future. As a project funded and supported by the Library of Congress the NGDA's mandate includes potential patrons who are substantially the same as those who

might use the Library of Congress itself. This does allow greater latitude in collecting as the potential patron base is so broad; anyone who might potentially benefit from these collections is included. While that might seem like carte blanche to collect broadly, both libraries follow their CDPs closely to ensure research level collections in their areas of focus. A number of elements set these digital geospatial collection policies apart from their analog brethren. In addition to sections on subject, scope, and area of coverage, digital geospatial CDPs also address metadata, format, versioning, proprietary formats, data set size, and ownership/access considerations.

Scope of Coverage

Much longer sections on the scope of coverage are included explicating how one would choose the type and scale of data to be collected. Suggestions are included about where to go to get the information once the types of data are chosen. For example, layers from the National Atlas and the National Map that are relevant for a particular institution are suggested. In the case of Stanford, with its strong focus on earth sciences, ten National Atlas layers were chosen for downloading focusing on volcanoes, historical earthquake data, bathymetry, geologic fault, and seismic hazard layers.

Metadata

Unlike analog materials, such as books, where an item's metadata travels with it in the form of title pages, or maps, where the margins contain legends, titles, publication data, and the like, digital data sets may not carry their own metadata. Such information often travels in an accompanying file. Ensuring that these data are captured is therefore critical to archiving digital datasets. Such metadata is vital for understanding the purpose of a particular data set, and includes when it was created, by whom it was created, technical information of interest to researchers such as the equipment (sensors, lens, settings) that created the data, and copyright information. Metadata regarding format may also be important for extracting data from an archive. Knowledge of the format in which the data were created is necessary for reconstituting it in order to view and utilize it, or migrate it to the next generation of software. Although there are standards for geospatial metadata creation, such as the Federal Geospatial Digital Clearinghouse (FGDC, 2008) standard, and ISO 19115:2003, not all data producers adhere to the recommended elements and the lack of uniform metadata is rampant. This lack of uniformity has ramifications for archiving geospatial data generally. Boxall (2004) noted the absence of standard procedures for archiving geospatial information and suggested that librarians work toward developing geospatial archiving standards. In 2008 standardized archiving procedures for geospatial data continue to be emerging rather than established.

Format

All digital data are created in a given format and this has implications for collecting and archiving. The CDPs for both UCSB and Stanford address the preference for obtaining data in open formats as well as the reality that a great deal of geospatial data are produced in proprietary formats. Thus both proprietary and open source formats are archived. It should be noted that neither node of the NGDA is archiving software.

Versioning

Analogous to paper map products, much digital geospatial data and imagery are produced at regular intervals to capture change over time. The collection policy addresses the need to collect at appropriate intervals to ensure holding both current and historical data. In the date/chronology section, the need for versioning is discussed as the decisions about when to collect content has shifted to the librarian and away from the content producers. Librarians and others charged with decision making must assess appropriate intervals for data collection. This will vary, of course, depending on the nature of the data being collected. The NGDA CDP describes issues that impact decisions regarding when and how often to collect including:

- new releases of data sets by the publisher when changes occur;
- trigger events such as the decennial census, or single events such as the aftermath of natural disasters;
- seasonal changes in the case of collecting Landsat scenes;
- changes in boundary lines such as school district or electoral districts;
- urban or infrastructure growth that impacts transportation routes or other observable features.

Proprietary Formats

Proprietary formats that are widely used are archived in both NGDA nodes. When ingested into the Stanford Digital Repository, the formats are described as fully as possible in a transfer manifest, essentially a digital packing slip that tells the repository what it is receiving.[1] Similarly, at UCSB proprietary formats are archived with a notation in the metadata that the format is proprietary. There is some concern over archiving proprietary formats. The central premise of the NGDA model for archiving is that format information is essential to the long-term value of any particular object. However, as proprietary formats are common for geospatial data, archives will have to make decisions about how to provide access to these data in the future. While the archives can preserve the data now for future users to view and utilize, such data may require that they have access to software to convert data from a previous format to the current format of the day. Format migration has been discussed by the nodes as a possible solution. It has not been tried on geospatial files and will be

considered on a case-by-case basis for each format in the future. Other agreements, yet to be negotiated with the creators of proprietary formats, could also enable the future use of such data.

Data Set Size and Ownership/Access Considerations

Geospatial data sets can be quite large. For some public domain data direct downloading is a possibility, however it is so labor intensive as to be prohibitive. As an example of data set size, the Stanford NGDA repository obtained data from the United States Geological Survey (USGS) for a relatively modest area of California and Nevada. Three sets of data, comprised of parts of the National Elevation Dataset (NED), the National Agriculture Imagery Program (NAIP) dataset, and high resolution orthoimagery for the Monterey Peninsula constituted 571 gigabytes of data. As public domain datasets, access considerations due to copyright or licensing restrictions are not a concern. The NGDA repositories are not intended to be dark archives. Although scenarios can be envisioned in which there would be a temporary moratorium on access, this would not be ideal. In any case, even if materials were archived but not available to the public, discovery of the materials would be provided for as well as information on how to obtain the data from the originating source.

One instance of ownership/access issues that the NGDA has encountered pertains to the California Geological Survey (CGS). There are serious access implications for many governmental agencies if they are funded according to how much the agency and its website are used. For example, the CGS has agreed to have the library at the University of California, Santa Barbara archive its digital data. However, the agreement stipulates that it is essential for the UCSB library to point its users of CGS data to the CGS website, to maximize that website's use. Direct access through UCSB will be allowed when the CGS website is not available.

Legal Agreements

Legal issues such as ownership and access, among others, needed to be addressed in order for the NGDA to acquire any collections that were *not* in the public domain. The NGDA lacked formal status independent of the universities sponsoring each node, therefore it was necessary to craft legal agreements that would satisfy the needs of depositors of content. Stanford had identified a significant private collection, the David Rumsey Map Collection, as its first goal of acquisition. Thus, even before the NGDA had completed an agreement between the two nodes governing their relationship, we created and finalized a Content Provider Agreement. Ongoing work is the completion of the NGDA Node Agreement (n.d.), which regulates how the nodes interact with each other. The challenge in creating a Node Agreement has been crafting an agreement whose policies and procedures work equally well for a public and private institution. The Con-

tent Provider Agreement and the accompanying Exhibit B document can be found at the NGDA website:

http://www.ngda.org/research/Rights/Stanford_NGDA_Contentprovider_102307final-1.pdf

http://www.ngda.org/research/Rights/NGDA_EXHIBITBcopy.pdf

Identification and Acquisition of Resources: Lessons of Acquisition

The ongoing work of the NGDA continues to yield valuable technical and nontechnical experience. The most significant areas where hands-on experience has resulted in new knowledge are metadata creation, acquisition of data, format issues, and storage of data.

Metadata Creation

An issue that stewards of digital collections grapple with is the variability of metadata. Not only is it the case that not all collections have good metadata, many have no metadata at all; and metadata creation, like the creation of standard catalog records, is an extremely time-consuming and therefore expensive activity. When faced with a situation where metadata is less than optimal choices include creating metadata where possible, choosing not to archive the data at all, or archiving with limited metadata while explaining to the depositor that less metadata may lead to less utility for the data in the future.

For example, the NGDA acquired a bundled variety of data and imagery accompanied by less than ideal metadata from a third party data aggregator. All of the data are in the public domain. The third party company added value to existing data by georectifying the images and digitizing features. The data included national shoreline and boundary data, military vector data, and Shuttle Radar Topography Mission data. The data was delivered with descriptive metadata and lacked any technical descriptions. After examining the company website and contacting customer service it was learned that all the metadata they had had been delivered. If possible, further metadata would need to be obtained from the government entities that originally created the data.

At times, metadata will be provided at both the collection level and at the individual level. For example, part of the NGDA content is the Landsat 7 images for the state of California and part of Arizona. This set (a subset of a worldwide collection of Landsat 7 imagery) contains thirty-eight scenes. Included with the metadata is information that is the same no matter which image is being viewed (such as the format, information about the satellite, the naming conventions for the scene and the bands, contact information, etc.). Also included with each image is the metadata related to each specific scene (such as the scene ID, the path and row numbers, and the date of acquisition). One then must decide how to store

the collection and individual metadata for each item. One then must decide how to store the collection level metadata and individual metadata for each item. Also, the collection level metadata and individual metadata are retrievable for each item in the collection.

One collection that has presented interesting metadata challenges is the David Rumsey Map Collection. Individual cataloging records have been created for nearly all of the items in the collection, but upon closer inspection, it was found that numerous images had no corresponding metadata, requiring individual reconciliation of each item with the best possible metadata record. This happened most frequently with atlases where there were images of the maps, the covers, the title page, and other text pages. The maps were automatically associated with the images by a corresponding file name and metadata field in the .xml record. (See appendix 2 for an example of a typical .xml record for a Rumsey collection image.) The extra imagery (covers, title pages, text) had no such corresponding file name in the associated metadata, requiring inspection of each item that failed to have matching fields. These records were inspected by hand and were associated to the main cataloging record that described the whole atlas. Even with excellent metadata records, the way in which the cataloging was done created idiosyncrasies when adding the images to the repository that did not allow for automated ingest of the whole collection.

Acquisition Issues for Compressed or Converted Files

Many geospatial data files are available for download in compressed formats. While this speeds download time and reduces storage requirements, for long-term archiving the native format is preferable. Storing files in their uncompressed original form is standard archiving protocol as compression can lead to bit loss. Compressed files also require another layer of quality control as they must be uncompressed to check that data are intact for archiving. While files being downloaded include geotifs, shapefiles, .dbf, and export files, many files being acquired for the archive have been converted into a transfer format. Numerous layers of the National Atlas, for example, are downloaded in a format called Spatial Data Transfer Standard (SDTS).[2] Similar to compressed files in that they are one-step removed from their original format, SDTS files may require validation prior to ingest into the repository.

Another step in readying data for ingest into an archive is creating a file structure and naming conventions to keep track of the data in whatever temporary staging location they are being held. The data and imagery files are checked for completeness and readability to ensure the files were not corrupted at some point in the process of creation or collection. Additional steps like these to verify the validity of datasets and organize them add labor time to the archive process and are an archiving cost that might not be obvious at the outset.

Format Issues

As discussed earlier, digital geospatial data are created in a number of different formats. Some geospatial formats, such as ESRI's shapefile, consist of multiple files that constitute one digital object and must travel together. This adds a layer of complexity when attempting to archive such files. This complexity was part of the impetus for creating a geospatial format registry. The goal of the format registry is to accurately and thoroughly describe as many of the formats as possible that are collected by the NGDA. Through description of the formats in a central location, all individual files can point to that information rather than having to pack it in its own metadata file. Registries of this sort typically include information about the nature of the object, the format creator, version information, and documentation written about the format. Understanding the formats in this way increases the likelihood of preservation in the future as they inevitably become obsolete.

The NGDA format registry is in its beginning stages with work presently focusing on the emerging standards across the field including comparison of models being set up by the Global Digital Format Registry (GDFR) and PRONOM. These registries have not been built out for geospatial formats hence the need to build out the existing models with relevant fields for these types of data and imagery. Five initial formats are being built out including ESRI's shapefiles, ESRI's ArcINFO GRIDs, .tif and geotifs, BIL and HDF files. Shapefiles are commonly used and examples from our collections include National Atlas and National Park Service layers as well as layers from the California Spatial Information Library (CASIL). Tif and geotif are image formats. Imagery of the California Bay Area produced by the National Agricultural Imagery Program (NAIP) is produced in geotif format. The Shuttle Radar Topography Mission (SRTM) topographic data are in .bil format. Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and Moderate Resolution Imaging Spectroradiometer (MODIS) imagery, often used for land cover analysis, is delivered in the .hdf format. It is hoped that by the end of the contract the registry will have information about all formats held by the NGDA.

Ingest and Short-Term Storage

Geospatial data that are selected for archiving go through a multistep process prior to ingest. After a collection is identified it may be downloaded from its original location, or received on storage media such as an external hard drive. It is then uploaded to a temporary storage location. At Stanford this is network attached storage (NAS). A protocol for incoming datasets has been developed. In order to track data from arrival, to storage, to archive, Stanford starts by creating a log, or manifest of the files upon receipt. An open source application called MD5 Deep is used. MD5 Deep walks through one directory at a time creating digests of each file found, and goes through

any subdirectories found. This "snapshot" of the data upon arrival is a safeguard that allows for detection of any future change or loss of files prior to ingest. The next step is mapping all descriptive metadata into MODS (Metadata Object Description Schema). Then a transfer manifest in .xml is created as described earlier, and this "packing slip" tells the repository what is being received. If no issues emerge during these stages, a sample item is ingested. If the test is successful, ingest of the full collection begins.

At UCSB the process is similar. The archive brings in data in a series of steps. Once a collection is identified as at-risk or otherwise desirable for preservation, the archive transfers it to local storage. Reliable transfer methods, such as Rsync (a robust internet file transfer tool), are used to ensure that the entirety of the dataset is received. In the case of extremely large datasets, or if a collection holder has an unreliable network connection, physical media may be transferred. Once the transfer has been completed, the data are ready to be worked with, but are not yet in the archive.

Before the data can be ingested into the archive, preliminary work must be done. Collections and sub-collections are identified, and when applicable, templates are created. Templates describe the structure of data within a (mostly) homogeneous collection. Once a collection is well understood and defined, the ingest processing begins. Configuration files are used to map disparate components into structured Archive Objects. Workflow tools allow for quality control prior to any interaction with the archive. When all checks are satisfied, Archive Objects are created, data are copied, and items are considered archived.

Conclusion

Collection development for a digital geospatial archive is complex, labor intensive, and challenging. Our approach has acknowledged the practical reality that we could not collect everything, nor would our institutions be able to justify the expense of collection building that was not aligned with the research goals of the institution. For an academic institution this model seems most likely to ensure success.

In the future the NGDA plans to expand by adding new nodes. One avenue being explored is to work through the Western Association of Map Libraries and the American Library Association roundtable group MAGERT (Map and Geography Round Table) to approach map libraries that are already collecting geospatial data. We believe as more institutions understand the need to archive geospatial materials, and also the challenges, that we will find a number of willing and able partners.

The NGDA nodes have worked to create a systematic process for selection, capture, and archiving. This has required close collaboration between the collections librarians and the digital archivists, including metadata specialists and programmers. The value of building such collections as we move forward in the digital age is significant and ensures

that part of the considerable geospatial data output of the nation is being captured for future use. Those responsible for collection development need to work closely with their technical counterparts in archive building to understand metadata requirements, short-term storage options, and any other factors that will impact the ultimate goal of acquiring geospatial data and imagery. The size of geospatial datasets will require cooperative strategies and federated archives among libraries and other archiving entities to ensure replication of materials. The NDIIPP program provided the necessary funding and impetus to initiate long-term preservation of geospatial materials at UCSB and Stanford, creating a partnership that will continue into the future and will pave the way for an expanding network of geospatial archives.

Notes

1. The transfer manifest is an *information package* per the description of the Open Archive Information System (OAIS) reference model released in May 1999. When information about the content object and its metadata are aggregated for purposes of ingestion into the Stanford Digital Repository, the transfer manifest or TM could be considered a Submission Information Package or SIP minus the actual inclusion of the content files themselves (Hoebelheinrich, 2008).
2. The USGS offers this definition of SDTS from its website:
   The Spatial Data Transfer Standard, or SDTS, is a robust way of transferring earth-referenced spatial data between dissimilar computer systems with the potential for no information loss. It is a transfer standard that embraces the philosophy of self-contained transfers, i.e. spatial data, attribute, georeferencing, data quality report, data dictionary, and other supporting metadata all included in the transfer. (U.S. Geological Survey, n.d.)

References

Ahonen-Rainio, P. (2005). Metadata for geographic information. *Journal of Map and Geography Libraries, 2*(1), 37–65.

Boxall, J. (2004). Advances and trends in geospatial information accessibility—Part I: Geolibraries. *Journal of Map and Geography Libraries, 1*(1), 7–39.

DePalma, A. (2008, September 15). Weather history offers insight into global warming. *New York Times*. Retrieved June 24, 2008, from http://www.nytimes.com/2008/09/16/science/earth/16moho.html?_r=1&oref=slogin

Federal Geographic Data Committee (2008). *Content standard for digital geospatial metadata (CSDGM) essential metadata elements*. Retrieved June 24, 2008 from http://www.fgdc.gov/metadata/documents/CSDGMEssentialMeta_20080514.pdf

Hoebelheinrich, N. (2008) *Metadata for Stanford Digital Repository*. Retrieved June 30, 2008, from https://consul.stanford.edu/display/DLMD/Metadata+for+Stanford+Digital+Repository

Leptoukh, G., Ahmad, S., Eaton, P., Koziana, J., Ouzounov, D., Savtchenko, A., Serafino, G., Sharma, A., Sikder, M.; Zhou, B. (July 2001). *MODIS data ingest, processing, archiving and distribution at the Goddard Earth Sciences DAAC*. Paper presented at the Geoscience and Remote Sensing Symposium, 2001. IGARSS 2001. IEEE 2001 International, Sydney, Australia.

National Geospatial Digital Archive (NGDA). (n.d.). NGDA content provider agreement. Retrieved June 30, 2008, from http://www.ngda.org/research/Rights/Stanford_NGDA_Contentprovider_102307final-1.pdf; http://www.ngda.org/research/Rights/NGDA_EXHIBITBcopy.pdf

Thiry, C. J. (Ed.). (2006). *Guide to U.S. map resources*. Lanham, MD: Scarecrow Press.

U.S. Geological Survey. (n.d.). What is SDTS? Retrieved May 21, 2008, from http://mcmcweb.er.usgs.gov/sdts/whatsdts.html

ATTACHMENT 1
COLLECTION DEVELOPMENT POLICY FOR THE NATIONAL GEOSPATIAL DIGITAL ARCHIVE
Final Version - November 1, 2006
http://www.ngda.org/research.php

I. Collection purpose and description of the users

The purposes of the NGDA collection are to archive broad collections of digital geospatial data of the United States, and make it available to users. The primary users of the NGDA are citizens—present and future—of the United States of America. The secondary users of NGDA are all other people who use the Web. Users of geospatial data are many; the following brief list is a sampling:

- From the university/academic world, undergraduate and graduate students and faculty, especially in those disciplines which deal with geographic areas, e.g., geography, anthropology, history, etc.
- From elementary and secondary schools, students and teachers looking for information about a specific country or city
- From the world of business and commerce, commercial vendors of imagery and mapping services; firms that need to know demographic information; realtors
- Non-profit, non-governmental organizations such as relief organizations; economic and social councils, etc.
- Persons and firms collecting information about the environment
- Government agencies - local, state, federal, international

II. Selection, Evaluation and Prioritization

Geospatial data are produced in large quantities from a wide-ranging group of organizations and entities. Data producers include government agencies, commercial vendors, and individuals. It is important to carefully evaluate the needs of the collecting institution in selecting materials to collect and archive. In order to begin the process, the following steps are recommended.

1. Develop awareness of potential collections.

- Contact and/or explore resources at local, regional, state, and federal agencies to find out what data are produced for the area of interest. Attend regional interest group meetings. Sign up for relevant e-lists such as GIS4lib and Maps-L. Read and subscribe to print and online publications focused on GIS data. Consider commercial sources for the area of interest.

2. Check that a potential collection is within the Scope of Coverage (see Section III below).

3. Assign a priority rating to each potential collection using the following questions as a starting point.

- Is collection in scope for the NGDA and for the NGDA collecting institution?
- Is the collection's geographic area of primary importance to the collecting institution?
- Is the collection at risk due to either:
  - The content provider does not archive the content.
  - The file format is becoming obsolete.

4. Obtain resources to collect first priority collections. Resources include collection funds if the data are not free, metadata/cataloging services, server or repository space, and resources to access the data such as computers and relevant software.

5. If resources are available, proceed to second priority collections.

III. Scope of coverage

The scope of collecting is solely in the realm of geospatial digital data. The term, "digital geospatial data," is defined as digital items, displayed as graphics, that are georeferenced or are geographically identified. These are primarily composed of: digital maps; remotely sensed images (e.g., aerial photographs; data collected by satellite sensors); datasets (e.g. shapefiles, layers, geodatabases, etc.); atlases; globes (celestial and terrestrial); aerial views (e.g., panoramas); block diagrams; geologic sections; topographic profiles; etc.

A. Geography

Primary collecting emphasis is on the geographic area of the United States. Data may cover the entire United States including U.S. territories; or data may be focused on a specific state, consortial area (e.g., southern California; metropolitan New York City), county, city, or city neighborhood.

i. United States national or large regional extent

Many subjects listed in Section B are available at a national level. For example, the United States Census Bureau publishes datasets for the geographic regions designated by their decennial census. National datasets available through numerous national government agencies and commercial entities cover many subjects and may be updated on a set schedule, such as census

data every ten years. Some of these datasets may be compiled in such a way to be of a manageable size for viewing, use, and preservation. Examples would include numerous layers in the National Atlas (http://www.nationalatlas.gov/). Others are very large and may require special treatment for archive consideration. For example, the United States Geological Survey produces and continues to update the National Elevation Dataset, served through the National Map (http://www.nationalmap.gov/) web site. The full dataset, as of May 2006, is sixty gigabytes of data. It may be the case at this point that such a dataset is too large for any one node to archive and so should be split across a number of archives. As of Spring 2006, many national agencies producing geospatial data are working out the policies for the archiving of their geospatial resources. Because policies are still being created, it is recommended that important datasets be archived at the local level to provide long-term access to the material.

ii. State, county, or city within the United States.

Datasets at this level are often produced by state, county, or city agencies. Coordination between groups often occurs because of the cost to produce some of the more expensive datasets. For example, this may be the case for aerial photography created at set intervals. Many states have created geospatial clearinghouses for dissemination of popular datasets. Preservation and retention of older datasets is not guaranteed nor often spelled out at the clearinghouse sites. Basemap data at each level are important to collect including transportation networks, boundary files, water resources, parcel information, and agency-specific data. Statewide clearinghouses are excellent sources for these data. Counties and cities may disseminate their data over the Internet, although it is highly likely one will need to contact these groups directly. Copyright status should be ascertained when collecting data at this level.

iii. Ocean-floor coverage: off-shore areas to the limit of the United States' maritime boundary claim.

Data of this type may include multi-beam surveys, sonar readings, electronic nautical charts, and vector data delineating boundary claims. It is created by national and state agencies, as well as academic institutions and research institutes.

B. Subject

Potential types of materials are split across physical and human/cultural geography. The list has been derived from the Library of Congress G Schedule and then modified.

i. Physical geography

- Mathematical geography (surveying and cartography; etc.)
- Physiography (e.g., topography)
- Hydrology
- Oceanography
- Rivers and lakes
- Floods
- Geology, geophysics, mineral resources, and soils
- Climatology
- Biogeography (land use/land cover)
- Flora
    - General
    - Aquatic
- Forests and forestry
- Agriculture
- Fauna
    - General
    - Aquatic

ii. Human and cultural geography

- Political geography
- Economics
- Real property; cadastre
- Public lands; ethnic reservations
- Demography; census
- Technology; engineering; public works
- Transportation and communication
- Commerce and trade; finance
- Military and naval geography
- Historical geography

iii. Remotely-sensed images

- Aerial photographs
- Satellite images

C. Date or Chronology

Geospatial data is often subject to versioning. New versions of data layers are released when changes occur in the original dataset. This can be done to correct errors or to account for changes over time. Data may be changed incrementally or on specific dates due to "trigger-events," such as the decennial census. Once it has been decided which datasets are important to archive, a decision should be made about the frequency with

which versioning should occur. For example, Landsat imagery might be collected once a quarter to follow seasonal changes. School district lines may only need to be captured when boundaries are re-drawn. Transportation routes could be updated on a yearly basis. Geospatial data collected immediately after a natural disaster would be collected as soon as it was made available to libraries and/or the public.

D. Format

As part of the NGDA project, only digital materials are collected. The digital data must be accompanied by minimum core required metadata. Section IV covers metadata recommendations. Open source, non-proprietary file formats that are either readily manipulated using standard image-processing or geographic-information-system software are preferred (e.g., geotiff, GML). Data in proprietary formats or data whose display is dependent upon proprietary software (e.g., ArcInfo Coverage or GRID) will be dealt with on a case-by-case basis. Some important factors will be how commonly available and used the software is and whether the data may be exported to a non-proprietary format. For example, the ESRI shapefile is a proprietary format, but it is so universally used, the current NGDA nodes will accept data in this format. Format registries for geospatial data are presently being created to capture relevant representational information about file formats. The Library of Congress has posted the "Sustainability of Digital Formats Planning for Library of Congress Collections" on their Web site (http://www.digitalpreservation.gov/formats/intro/intro.shtml). At present, it contains no information about geospatial file formats, but certainly will in the future. The Global Digital Format Registry (GDFR) is also interested in including geospatial format information (http://hul.harvard.edu/gdfr/). Appendix 2 includes definitions of some of the more common geospatial data formats. File formats will change over time with new formats being created and older formats falling into disuse. The Archives should continually evaluate if it is possible to migrate older formats into newer ones, decide when and if old formats will be kept, and keep abreast of best practices in the geospatial community regarding file formats.

E. Language

English will be the most frequent language collected due to the nature and focus of the grant; but it is possible that geographic coverage for areas in the U.S. may have text portions in languages other than English.

F. Copyright

Materials without copyright—e.g., public domain data—comprise much of the collection. Acquisition and access to copyrighted data will be gov-

erned by an agreement between the NGDA collection node and the data provider. Access to certain data may be restricted for a specified period of time at the request of the data provider. The NGDA nodes do not intend to be dark archives.

G. Exclusions:

- Digital information with a geographic reference, such as a text history of Alabama
- Analog materials
- Straight statistical data not tied to a geographic area
- Data layers external to the United States
- Data that is to remain perpetually "dark"
- The HTML content from Web sites (geospatial data gathered from Web sites will be collected)

IV. Metadata recommendations

It is important to collect as much metadata as is feasible. The NGDA recommends a core set of metadata fields with the understanding that it may not be available in all cases. Whether or not content lacking core fields will be ingested into a node is up to the node itself. Significant metadata in rough order of importance:

- Geographic area: This includes information about the extent of the content. Specific node requirements could require coordinates in decimal degrees or words describing the extent ("Arizona counties" qualified by years).
- Type (intellectual content): This includes maps, remote-sensing imagery (aerial photograph; image from satellite), layers.
- Format: This should identify the file types included (e.g., tiff, jpeg, arcexport, shapefile).
- Projection and/or coordinate system
- Scale and/or resolution: Resolution is often cited when dealing with aerial and satellite imagery.
- Transfer media: This component details the device upon which the data are stored when deposited with the archive (CD-ROM, DVD-ROM, hard-drive, etc.).
- Title: A title is required for each item ingested. A title may also be included for the whole collection.
- Date of information: This would be the date the information was created.
- Issuance information: This includes the issuing agency, the place of issuance, and date of issuance.
- Data Quality information e.g. FGDC metadata elements such as attribute accuracy and completeness report.

- Rights Information e.g. copyright, reproduction of data
- Date ingested into the archive
- Contact information for the content provider: This would potentially include a contact person(s), address, telephone/fax, email addresses, or Web site.
- Collection name and description: This may be supplied by the node ingesting the data.
- Controlled-list subject headings: This might be created by the content provider. Hopefully they would provide a full copy, but at the very least a full "bibliographic" citation.
- Other fields: If content provider provides any fields, then they should also include the field name, field definition, and domain (an authority list).

V. Sources for digital geospatial data

Although governmental sources of data are of primary interest, the archive is not to be limited to data generated, or contracted, by federal agencies of the United States, but instead will include digital geospatial data generated by any agency or person. The emphasis will be firstly on nationwide coverage, of any theme, and very often these are generated by federal agencies. It is expected that nodes will collect in this area according to their specific regional and research needs. This may mean that part of the collecting decision is made by the scale of the dataset.The NGDA nodes will focus on government agencies at all levels and non-profit entities such as professional organizations or environmentally focused non-profits, with a secondary focus on commercial firms, and a tertiary focus on products issued by people.

VI. Coordination and cooperation with other collections

The NGDA's mission is to be a collecting network. As such, collaboration with other institutions is expected and necessary. Because digital geospatial data sets require large amounts of server space, the cooperation of many institutions will be necessary to build an extensive collection. Cooperation agreements written specifically to govern the collecting areas of the partners should include:

- the collecting areas for each participating institution;
- the frequency of updates and versioning for each dataset;
- the length of the agreement;
- the type and level of access to be provided to the collected materials;
- a set interval to review the collection agreement;
- a glossary.

VII. Appendices

Appendix 1:

Sample NGDA-node collection development policies

Note: Once this policy has been vetted and finalized, both UC Santa Barbara and Stanford will write Collection Development Policies for their nodes.

Appendix 2: Glossary

Digital Elevation Model

Digital Elevation Models display a three-dimension-like image of surface elevation by using a raster grid of evenly spaced elevation values. The values are obtained from USGS topographic maps.
Compiled from multiple sources including:
Retrieved June 10, 2006, http://edc.usgs.gov/products/elevation/dem.html
Retrieved June 10, 2006, http://www.landinfo.com/resources_dictionaryAD.htm#d

Digital Orthophoto Quadrangle
A digital orthophoto quadrangle (DOQ) is a computer-generated georeferenced image of an aerial photograph in which image displacement caused by terrain relief and camera tilts has been removed. It combines the image characteristics of a photograph with the geometric qualities of a map.
Retrieved June 9, 2006, http://www.usgsquads.com/prod_doqq.htm

Digital Raster Graphic
A digital raster graphic (DRG) is a scanned image of a U.S. Geological Survey (USGS) standard series topographic map, including all map collar information. The image inside the map neatline is georeferenced to the surface of the earth and fit to the Universal Transverse Mercator projection. The horizontal positional accuracy and datum of the DRG matches the accuracy and datum of the source map. The map is scanned at a minimum resolution of 250 dots per inch.
Retrieved June 9, 2006, http://topomaps.usgs.gov/drg/

GML
GML is an acronym for Geography Markup Language. An OpenGIS Implementation Specification designed to store and transport geographic information. GML is a profile (encoding) of XML.

Compiled from multiple sources including:
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=gml

Georeferencing
To establish a relationship between page coordinates on a planar map and known real-world coordinates. Georeferencing allows geographic data sets to be analyzed and compared with one another.
Compiled from multiple sources including:
Retrieved July 31, 2006, http://www.geo.ed.ac.uk/agidexe/term?1228
Retrieved July 6, 2006, http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=georeference

Geospatial
Relating to physical features of the earth and their geographic location, including both natural and man-made features. Geospatial data refers to information derived from maps or remote sensing techniques, such as aerial photography or satellite imagery.
Compiled from multiple sources including:
Webster's New Millennium Dictionary of English, Preview Edition (v 0.9.6)
Copyright © 2003-2005 Lexico Publishing Group, LLC
Retrieved June 8, 2006, http://dictionary.reference.com/search?q=geospatial&r=66
Directions Magazine: the worldwide source for Geospatial Technology
Retrieved July 26, 2006, http://www.directionsmag.com/press.releases/index.php?duty=Show&id=10412&trv=1

Raster
An image formed using individual dots with color values, called cells (or pixels). Cells are viewed in a rectangular grid with each cell evenly spaced. Aerial photographs and satellite images are examples of raster images used in mapping.
Raster layers in a GIS system can depict such information as elevation, precipitation, and temperature.
Compiled from multiple sources including:
Retrieved July 28, 2006, http://data.geocomm.com/helpdesk/glossary-r.html

Remote Sensing
RS is the process of using a recording device not in physical contact with the surface being analyzed to obtain data.

Methods include aerial photography and using sensors sensitive to various bands of the electromagnetic spectrum. Equipment can be deployed from aircraft, satellite or space probe.
Compiled from multiple sources including:
Retrieved July 28, 2006, http://data.geocomm.com/helpdesk/glossary-r.html

Shapefile
Shapefile is the name of the proprietary digital vector storage format created by ESRI Corporation. Shapefiles are used and created in software such as ArcView, Arc/Info, ArcGIS and other widely used GIS software. A shapefile consists of multiple files that together generate a data layer in a Geographic Information System (GIS). There are three required files that are stored and deployed together in a shapefile:

.shp - the file that stores the feature geometry.
.shx - the file that stores the index of the feature geometry.
.dbf - the dBASE file that stores the attribute information of features.

Other files can be added to the shapefile to carry additional information such as projection and metadata.
Compiled from multiple sources including:
Retrieved June 29, 2006, http://en.wikipedia.org/wiki/ESRI_shapefiles
Retrieved June 29, 2006, http://walrus.wr.usgs.gov/infobank/programs/html/definition/shapefile.html
Retrieved July 26, 2006, http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=shapefile

Vector
Vector data (used in a GIS system) is one method used to store spatial data. Features are defined by their boundaries only and curved lines are represented as a series of connecting arcs. Vector data is expressed as X,Y,Z coordinates. Examples of vector layers include schools (points), street networks (lines), and voting districts (polygons).
Compiled from multiple sources including:
Retrieved June 9, 2006, http://www.geo.ed.ac.uk/agidexe/term?349
Retrieved July 31, 2006, http://data.geocomm.com/helpdesk/glossary-v.html

XML
XML is an acronym for Extensible Markup Language. Developed by the World Wide Web Consortium (W3C), it is a standardized markup language for designing text formats. It enables the interchange of data be-

tween computer applications. XML is a set of rules for creating standard information formats using customized tags and sharing both the format and the data across applications.
Compiled from multiple sources including:
http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.search&search=true&searchTerm=xml GIS Dictionaries and Glossaries:
http://www.agi.org.uk/bfora/systems/xmlviewer/default.asp?arg=DS_AGI_TRAINART_67/_firsttitle.xsl/87http://www.fgdc.gov/metadata/csdgm/glossary.htmlhttp://www.gis.com/whatisgis/glossaries.htmlhttp://www.landinfo.com/resources_dictionaryAD.htm

Appendix 3:

Collection Levels

0 - Out of Scope
1 - Minimal Level
2 - Basic Level
3 - Study Level
4 - Research Level
5 - Comprehensive Level

Appendix 4: For more information (links and bibliography)

- The Center for International Earth Science Information Network (CIESEN)
  http://www.ciesin.org/
- Digital Curation Centre, United Kingdom
  http://www.dcc.ac.uk/
- Federal Geographic Data Committee (FGDC)
  http://www.fgdc.gov/
- FGDC Content Standard
  http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html
- National Archives and Records Administration (NARA)
  Requirements for transfer of permanent electronic records, geospatial data: http://www.archives.gov/records-mgmt/initiatives/digital-geospatial-data-records.html
- National Digital Information Infrastructure and Preservation Program (NDIIPP)
  http://www.digitalpreservation.gov/
- National Geospatial Digital Archive (NGDA)
  http://www.ngda.org

- National Library of Australia, Preserving Access to Digital Information (PADI)
  http://www.nla.gov.au/padi/index.html
- North Carolina Geospatial Data Archiving Project (NDGDAP)
  http://www.lib.ncsu.edu/ncgdap/
- Maine GeoArchives: A collaborative project between the Maine State Archives and the Geolibrary Board
  http://www.maine.gov/sos/arc/GeoArchives/geoarch.html
- United States National Satellite Land Remote Sensing Data Archive
  http://edc.usgs.gov/archive/nslrsda/

Tracey Erwin is a geospatial librarian for the National Geospatial Digital Archive (NGDA), a grant-funded project of the University of California, Santa Barbara and Stanford University. The NGDA is a collaborative initiative funded by the Library of Congress to identify, collect, and preserve geospatial digital materials within a nationwide digital preservation infrastructure. Tracey has worked on collection development, selection, and acquisition of data and imagery, and legal agreements for the Archive since 2005.

Julie Sweetkind-Singer is the head librarian at the Branner Earth Sciences Library and Map Collections at Stanford University. Her subject specialization is maps and GIS. She is currently Stanford's project lead on an NDIIPP grant from the Library of Congress. She has worked at Stanford since May 2000. In 1999 she worked jointly with David Rumsey on the Rumsey Map Collection Web site, which displays over 12,000 maps from the eighteenth and nineteenth centuries. She was the president of the Western Association of Map Libraries from June 2004 to July 2005. She was the vice-president of the California Map Society, Northern Chapter, in 2001 and 2002. She received her MLIS from San Jose State University and MBA from the University of Colorado at Boulder.

Mary Lynette Larsgaard is the director of the Davidson Map Library at the University of California, Santa Barbara. She has also been assistant head of the Map and Imagery Laboratory, Davidson Library, since 1988. The Map and Imagery Lab has a collection of remote-sensing imagery and maps of approximately 5.5 million items and is the largest of its kind in any university library in North America. Larsgaard has published extensively in the field of geospatial data in libraries, most notably with a widely used text, *Map Librarianship: An Introduction* (now in a third edition, published in 1998 by Libraries Unlimited); she is also a coeditor (with Paige Andrew) of the *Journal of Map and Geography Libraries/Geoscapes*. Her specialties are cataloging/metadata creation and twentieth-century and more recent topographic and geologic maps. In the year 2000 she was promoted to librarian, distinguished step, a promotion given only to librarians who have demonstrated superior competence and are internationally recognized as an authority in an area of library science.